# A APPENDIX

# B NUMERICAL RESULTS PARAMETERS

All experiments were done on a computer with 32gb of RAM, a CPU Intel Core i9-9900K @3.60GHz x 16, a GPU GeForce RTX 2080 Ti/PCIe/SSE2, and Ubuntu $18.04.4$ LTS.

## B.1 RECOMMENDATION SYSTEM

We split the dataset with $90\%$ for the training set, and $10\%$ for the testing set, and we run $20$ independent random partitions. For the optimizer, we used $5$ samples for the batch size, and ADAM algorithm Kingma & Ba (2015), with learning rate $0.005$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and without learning rate decay. For the loss, we used the smooth $L_1$ loss. For the GNN, we used ReLU as non-linearity, we considered $F = 32$ features, $K = 5$ filter taps, and $L = 1$ layers.

We used the graph neural networks library available online at https://github.com/alelab-upenn/graph-neural-networks/blob/master/examples/movieGNN.py and implemented with PyTorch.

## B.2 DECENTRALIZED CONTROL

We run the system for $T = 2s$, and used $400$ samples for training, $20$ for validation, and $20$ for the test set. For the optimizer, we used $20$ samples for the batch size, and ADAM algorithm Kingma & Ba (2015) with learning rate $0.0005$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, without learning rate decay. We used a one layer Graph Neural Networks with $F = 64$ hidden units and $K = 3$ filter taps, and used the hyperbolic tangent as non-linearity $\rho$. We run $10$ independent realizations of each experiment.

We used the graph neural networks library available online at https://github.com/alelab-upenn/graph-neural-networks/blob/master/examples/flockingGNN.py and implemented with PyTorch.

## C PROOF OF THEOREM 1

**Definition 2** (Template graphs)**.** *Let $\{u_i\}_{i=1}^n$ be the regular $n$-partition of $[0,1]$, i.e.,*

$$u_i = \frac{i-1}{n} \tag{19}$$

*for $1 \leq i \leq n$. The $n$-node template graph $\mathbb{G}_n$, whose GSO we denote $\mathbb{S}_n$, is obtained from $\mathbf{W}$ as*

$$[\mathbb{S}_n]_{ij} = \mathbf{W}(u_i, u_j) \tag{20}$$

*for $1 \leq i, j \leq n$.*

**Definition 3** (Graphon spectral representation of convolutional filter response)**.** *As the graphon $\mathbf{W}$ is bounded and symmetric, $T_{\mathbf{W}}$ is a self adjoint Hilbert-Schmidt operator, which allows to use the operator's spectral basis $\mathbf{W}(u,v) = \sum_{i \in \mathbf{Z}\{0\}} \lambda_i \psi_i(u) \psi_i(v)$. Eigenvalues $\lambda_i$ are ordered in decreasing order of absolute value i.e., $1 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq 0 \geq \cdots \geq \lambda_{-2} \geq \lambda_{-1} \geq -1$, and their only accumulation point is $0$ (Lax, 2002, Theorem 3, Chapter 28). Thus, we define the spectral representation of the convolutional filter $T_{\mathbf{H}}$ (cf.* (6)*) as,*

$$h(\lambda) = \sum_{k=0}^{K-1} h_k \lambda^k \tag{21}$$

**Definition 4** ($c$-band cardinality of $\mathbf{W}$)**.** *The $c$-band cardinality, denoted $B_{\mathbf{W}}^c$, is the number of eigenvalues whose absolute value is larger than $c$.*

$$B_{\mathbf{W}}^c = \#\{\lambda_i : \|\lambda_i\| \leq c\} \tag{22}$$

**Definition 5** ($c$-eigenvalue margin of $\mathbf{W}$ - $\mathbf{W}_n$)**.** *The $c$-eigenvalue margin of $\mathbf{W}$ - $\mathbf{W}_n$ is defined as the minimum distance between two different eigenvalues of the integral operator applied to $\mathbf{W}$, and to $\mathbf{W}_n$ as follows,*

$$\delta_{\mathbf{W}\mathbf{W}_n}^c = \min_{i, j \neq i} \{\|\lambda_i(T_{\mathbf{W}}) - \lambda_i(T_{\mathbf{W}_n})\| : \|\lambda_i(T_{\mathbf{W}_n})\| \geq c\} \tag{23}$$

**Definition 6** (Graphon Convolutional Filter)**.** *Given a graphon $\mathbf{W}$, a graphon signal $X$, and filter coefficients $\mathbf{h} = [h_0, \ldots, h_{K-1}]$ the graphon filter $T_{\mathbf{H}} : L_2([0,1]) \to L_2([0,1])$ is defined as,*

$$(T_{\mathbf{H}} X)(v) = \sum_{k=0}^{K-1} h_k (T_{\mathbf{W}}^{(k)} X)(v). \tag{24}$$

**Proposition 1.** *Let $X \in L_2([0,1])$ be a normalized Lipschitz graphon signal, and let $X_n$ be the graphon signal induced by the graph signal $\mathbf{x}_n$ obtained from $X$ on the template graph $\mathbb{G}_n$ [cf. Definition 2], i.e., $[\mathbf{x}_n]_i = X((i-1)/n)$ for $1 \leq i \leq n$. It holds that*

$$\|X - X_n\|_{L_2} \leq \frac{1}{n}. \tag{25}$$

*Proof.* Let $I_i = [(i-1)/n, i/n)$ for $1 \leq i \leq n-1$ and $I_n = [(n-1)/n, 1]$. Since the graphon is normalized Lipschitz, for any $u \in I_i$, $1 \leq i \leq n$, we have

$$\|X(u) - X_n(u)\| \leq \max\left(\left|u - \frac{i-1}{n}\right|, \left|\frac{i}{n} - u\right|\right) \leq \frac{1}{n}. \tag{26}$$

We can then write

$$\|X - X_n\|^2 = \int_0^1 |X(u) - X_n(u)|^2 du \tag{27}$$

$$\leq \int_0^1 \left(\frac{1}{n}\right)^2 du = \left(\frac{1}{n}\right)^2, \tag{28}$$

which completes the proof. □

**Proposition 2.** *Let $\mathbf{W} : [0,1]^2 \to [0,1]$ be a normalized Lipschitz graphon, and let $\mathbb{W}_n := \mathbf{W}_{\mathbb{G}_n}$ be the graphon induced by the template graph $\mathbb{G}_n$ generated from $\mathbf{W}$ as in Definition 2. It holds that*

$$\|\mathbf{W} - \mathbb{W}_n\| \leq \frac{2}{n}. \tag{29}$$

*Proof.* Let $I_i = [(i-1)/n, i/n)$ for $1 \leq i \leq n-1$ and $I_n = [(n-1)/n, 1]$. Since the graphon is Lipschitz, for any $u \in I_i$, $v \in I_j$, $1 \leq i, j \leq n$, we have

$$\|\mathbf{W}(u,v) - \mathbb{W}_n(u,v)\| \leq \max\left(\left|u - \frac{i-1}{n}\right|, \left|\frac{i}{n} - u\right|\right) \tag{30}$$

$$+ \max\left(\left|v - \frac{j-1}{n}\right|, \left|\frac{j}{n} - v\right|\right) \tag{31}$$

$$\leq \frac{1}{n} + \frac{1}{n} = \frac{2}{n}. \tag{32}$$

We can then write

$$\|\mathbf{W} - \mathbb{W}_n\|^2 = \int_0^1 |\mathbf{W}(u,v) - \mathbb{W}_n(u,v)|^2 du dv \tag{33}$$

$$\leq \int_0^1 \left(\frac{2}{n}\right)^2 du dv = \left(\frac{2}{n}\right)^2 \tag{34}$$

which concludes the proof. □

**Proposition 3.** *Consider the $L$-layer WNN given by $Y = \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})$, where $F_0 = F_L = 1$ and $F_\ell = F$ for $1 \leq \ell \leq L-1$. Let $c \in (0,1]$ and assume that the graphon convolutions in all layers of this WNN have $K$ filter taps [cf. (6)]. Under Assumptions 1 through 3, the norm of the gradient of the WNN with respect to its parameters $\mathcal{H} = \{\mathbf{H}_{lk}\}_{l,k}$ can be upper bounded by,*

$$\|\nabla_{\mathcal{H}} \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})\| \leq F^{2L}\sqrt{K}. \tag{35}$$

*Proof.* We will find an upper bound for any element $[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}$ of the tensor $\mathcal{H}$. We start by the last layer of the WNN, applying the definition given in equation (8),

$$\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})\| = \left\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X_L^f\right\| \tag{36}$$

$$= \left\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \rho\left(\sum_{g=1}^{F_{l-1}} \sum_{k=1}^{K-1} (T_{\mathbf{W}}^{(k)} X_{l-1}^g)[\mathbf{H}_{Lk}]_{gf}\right)\right\|. \tag{37}$$

By Assumption 3, the non-linearity $\rho$ is normalized Lipschitz , i.e. $\nabla\rho(\cdot)(u) \leq 1$ for all $u$. Thus, aplying the chain rule for the derivative, and the Cauchy-Schwartz inequality, the right hand side of the previous expression can be rewritten as,

$$\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})\| = \left\|\nabla\rho\left(\sum_{g=1}^{F_{l-1}} \sum_{k=1}^{K-1} (T_{\mathbf{W}}^{(k)} X_{l-1}^g)[\mathbf{H}_{Lk}]_{gf}\right)\right\|$$

$$\left\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \sum_{g=1}^{F_{l-1}} \sum_{k=1}^{K-1} (T_{\mathbf{W}}^{(k)} X_{l-1}^g)[\mathbf{H}_{Lk}]_{gf}\right\| \tag{38}$$

$$\leq \left\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \sum_{g=1}^{F_{l-1}} \sum_{k=1}^{K-1} (T_{\mathbf{W}}^{(k)} X_{l-1}^g)[\mathbf{H}_{Lk}]_{gf}\right\| \tag{39}$$

Note that the a larger bound will occur if $l^\dagger < L-1$, then by linearity of derivation, and the triangle inequality we obtain,

$$\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})\| \leq \sum_{g=1}^{F_{l-1}} \left\| \sum_{k=1}^{K-1} T_{\mathbf{W}}^{(k)} (\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X_{l-1}^g)[\mathbf{H}_{Lk}]_{gf}\right\| \tag{40}$$

By Assumption 2, the convolutional filters are non-amplifying, thus it holds that,

$$\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \mathbf{\Phi}(X;\mathcal{H},\mathbf{W})\| \leq \sum_{g=1}^{F_{l-1}} \left\| \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X_{l-1}^g \right\| \tag{41}$$

Now note that as filters are non-amplifying, the maximum difference in the gradient will be attained at the first layer ($l = 1$) of the WNN. Also note that the derivative of a convolutional filter $T_{\mathbf{H}}$ [cf. Definition 6] at coefficient $k^\dagger = i$, is itself a convolutional filter with coefficients $\mathbf{h}_i$. The values of $\mathbf{h}_i$ are $[\mathbf{h}_i]_j = 1$ if $j = i$ and 0 otherwise. Thence,

$$\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \mathbf{\Phi}(X;\mathcal{H},\mathbf{W})\| \leq F^{L-1} \left\| \mathbf{h}_{i *_{\mathbf{W}}} X_0 \right\| \tag{42}$$

$$\leq F^{L-1} \|X_0\|. \tag{43}$$

To complete the proof note that tensor $\mathcal{H}$ has $F^{L-1}K$ elements, and each individual gradient is upper bounded by (43), and $\|X\|$ is normalized by Assumption 1. □

**Lemma 1.** *Let $\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})$ be a WNN with $F_0 = F_L = 1$, and $F_l = F$ for $1 \leq l \leq L-1$. Let $c \in (0,1]$, and assume that the graphon convolutions in all layers of this WNN have $K$ filter taps [cf. (6)]. Let $\mathbf{\Phi}(\mathbf{x}_n;\mathcal{H},\mathbf{S}_n)$ be a GNN sampled from $\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})$ as in (9). Under assumptions (1),(2),(3), and (5) with probability $1-\xi$ it holds that,*

$$\|\mathbf{\Phi}(X;\mathcal{H},\mathbf{W}) - \mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n)\| \leq LF^{L-1}\left(1 + \frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}}\right) \frac{2\left(1+\sqrt{n\log(\frac{2n}{\xi})}\right)}{n} + \frac{1}{n} + 4LF^{L-1}c \tag{44}$$

*The fixed constants $B^c_{\mathbf{W}}$ and $\delta^c_{\mathbf{W}\mathbf{W}_n}$ are the $c$-band cardinality and the $c$-eigenvalue margin of $\mathbf{W}$ and $\mathbf{W}_n$ respectively [cf. Definitions 4,5].*

*Proof.* We start by writing the expression on the left hand side, using the definition of WNN [cf. (8)] we can write,

$$\|\mathbf{\Phi}(X;\mathcal{H},\mathbf{W}) - \mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n)\| = \|X_L - X_{nL}\| \tag{45}$$

$$= \left\| \rho\left( \sum_{g=1}^{F_{L-1}} \sum_{k=1}^{K-1} (T^{(k)}_{\mathbf{W}} X^g_{L-1})[\mathbf{H}_{Lk}]_{gf} \right) - \rho\left( \sum_{g=1}^{F_{L-1}} \sum_{k=1}^{K-1} (T^{(k)}_{\mathbf{W}_n} X^g_{nL-1})[\mathbf{H}_{Lk}]_{gf} \right) \right\|.$$

Since the non-linearity $\rho$ is normalized Lipschitz by Assumption 3, using the triangle inequality, we obtain

$$\|X_L - X_{nL}\| \leq \sum_{g=1}^{F_{L-1}} \left\| \sum_{k=1}^{K-1} (T^{(k)}_{\mathbf{W}} X^g_{L-1})[\mathbf{H}_{Lk}]_{gf} - \sum_{k=1}^{K-1} (T^{(k)}_{\mathbf{W}_n} X^g_{nL-1})[\mathbf{H}_{Lk}]_{gf} \right\|. \tag{46}$$

Using the triangle inequality once again, we split the last inequality into two terms as follows,

$$\|X_L - X_{nL}\| \leq \sum_{g=1}^{F_{L-1}} \left\| \sum_{k=1}^{K-1} T^{(k)}_{\mathbf{W}} (X^g_{L-1} - X^g_{nL-1})[\mathbf{H}_{Lk}]_{gf} \right\| \textbf{ (1)}$$

$$+ \sum_{g=1}^{F_{L-1}} \left\| \sum_{k=1}^{K-1} (T^{(k)}_{\mathbf{W}} - T^{(k)}_{\mathbf{W}_n}) X^g_{L-1}[\mathbf{H}_{Lk}]_{gf} \right\| \textbf{ (2)}. \tag{47}$$

Where we have split (47) into terms **(1)**, and **(2)**. On the one hand, by assumption 2, convolutional filters $h$ are non-amplifying, thus using Cauchy-Schwartz inequality, term **(1)** can be bounded by,

$$\sum_{g=1}^{F_{L-1}} \left\| \sum_{k=1}^{K-1} T^{(k)}_{\mathbf{W}} (X^g_{L-1} - X^g_{nL-1})[\mathbf{H}_{Lk}]_{gf} \right\| \leq \sum_{g=1}^{F_{L-1}} \left\| X^g_{L-1} - X^g_{nL-1} \right\|. \tag{48}$$

To bound term **(2)**, denoting $h_{Lgf}$ the spectral representation of the convolutional filter applied to $X^g_{L-1}$ at feature $f$ of layer $L$ [cf. Definition 3], we will decompose the filter as follows,

$$h^{\geq c}_{Lgf}(\lambda) \begin{cases} 0 & \text{if } |\lambda| < c \\ h_{Lgf}(\lambda) - h_{Lgf}(c) & \text{if } |\lambda| \geq c \end{cases} \tag{49}$$

$$h^{<c}_{Lgf}(\lambda) \begin{cases} h_{Lgf}(\lambda) & \text{if } |\lambda| < c \\ h_{Lgf}(c) & \text{if } |\lambda| \geq c. \end{cases} \tag{50}$$

Note that $h_{Lgf} = h^{\geq c}_{Lgf} + h^{<c}_{Lgf}$. Let $T^{<c}_{[\mathbf{H}_L]_{gf}}$ and $T^{<c}_{[\mathbf{H}_{nL}]_{gf}}$, be the graphon convoutional filters with filter function $h^{<c}_{Lgf}$ on graphons $\mathbf{W}$, and $\mathbf{W}_n$ respectively [cf. Definition 6]. Note that filter $h^{<c}_{Lgf}$, varies only in the interval $[0, c)$, and since filters are normalized Lipschitz by Assumption 2, it verifies

$$\left\| T^{<c}_{[\mathbf{H}_L]_{gf}} X^g_{L-1} - T^{<c}_{[\mathbf{H}_{nL}]_{gf}} X^g_{L-1} \right\| \leq \left\| (h_{Lgf}(c) + c) - (h_{Lgf}(c) - c) \right\| \left\| X^g_{L-1} \right\| \tag{51}$$

$$\leq 2c \| X^g_{L-1} \|. \tag{52}$$

Now we need to upper bound the difference in the high frequencies $h^{\geq c}_{Lgf}$. Let $T^{\geq c}_{[\mathbf{H}_L]_{gf}}$ and $T^{\geq c}_{[\mathbf{H}_{nL}]_{gf}}$, be the graphon filters with filter function $h^{\geq c}_{Lgf}$ on graphons $\mathbf{W}$, and $\mathbf{W}_n$ respectively. Let $\mathbb{S}_n$ denote the template graph sampled from the graphon $\mathbf{W}$ [cf. definition 2]. We denote $\mathbb{W}_n$, the induced graphon by template graph $\mathbb{S}_n$ as in (10). By introducing $T^{\geq c}_{[\mathbb{H}_{nL}]_{gf}}$, the graph filter with filter function $h^{\geq c}_{Lgf}$ on graphon $\mathbb{W}_n$, we can use the triangle inequality to obtain,

$$\begin{aligned} \left\| T^{\geq c}_{[\mathbf{H}_L]_{gf}} X^g_{L-1} - T^{\geq c}_{[\mathbf{H}_{nL}]_{gf}} X^g_{L-1} \right\| \leq & \left\| T^{\geq c}_{[\mathbf{H}_L]_{gf}} X^g_{L-1} - T^{\geq c}_{[\mathbb{H}_{nL}]_{gf}} X^g_{L-1} \right\| \textbf{(2.1)} \\ & + \left\| T^{\geq c}_{[\mathbb{H}_{nL}]_{gf}} X^g_{L-1} - T^{\geq c}_{[\mathbf{H}_{nL}]_{gf}} X^g_{L-1} \right\| \textbf{(2.2)}. \end{aligned} \tag{53}$$

Under assumptions 1–5, to bound term **(2.1)** we can use (Ruiz et al., 2020a, Theorem 1), and to bound term **(2.2)** we can use (Ruiz et al., 2020d, Lemma 2). Thus, with probability $1 - \xi$, the previous expression can be bounded by,

$$\left\| T^{\geq c}_{[\mathbf{H}_L]_{gf}} X^g_{L-1} - T^{\geq c}_{[\mathbf{H}_{nL}]_{gf}} X^g_{L-1} \right\| \leq \left( 1 + \frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}} \right) \frac{2\left( 1 + \sqrt{n \log(\frac{2n}{\xi})} \right)}{n} \left\| X^g_{L-1} \right\|. \tag{54}$$

Where the fixed constants $B^c_{\mathbf{W}}$ and $\delta^c_{\mathbf{W}\mathbf{W}_n}$ are the $c$-band cardinality and the $c$-eigenvalue margin of $\mathbf{W}$ and $\mathbf{W}_n$ respectively [cf. Definitions 4,5]. Hence, coming back to (47), we can use (48) to upper bound **(1)**, and we can use (52), and (54), to upper bound **(2)** as follows,

$$\begin{aligned} \| X_L - X_{nL} \| \leq & \sum_{g=1}^{F_{L-1}} \left\| X^g_{L-1} - X^g_{nL-1} \right\| + 2c \| X^g_{L-1} \| \\ & + \left( 1 + \frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}} \right) \frac{2\left( 1 + \sqrt{n \log(\frac{2n}{\xi})} \right)}{n} \left\| X^g_{L-1} \right\|. \end{aligned} \tag{55}$$

Now, we arrive at a recursive equation that we can compute for the $L$ layers, with $F$ features per layer, to obtain,

$$\begin{aligned} \| X_L - X_{nL} \| \leq & F_0 \| X_0 - X_{n0} \| + 2LF^{L-1} c \| X_0 \| \\ & + LF^{L-1} \left( 1 + \frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}} \right) \frac{2\left( 1 + \sqrt{n \log(\frac{2n}{\xi})} \right)}{n} \| X_0 \|. \end{aligned} \tag{56}$$

Using Proposition 1, noting that $F_0 = 1$ by construction, and using Assumption 1, concludes the proof. □

**Lemma 2.** *Let $\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})$ be a WNN with $F_0 = F_L = 1$, and $F_l = F$ for $1 \leq l \leq L-1$. Let $c \in (0,1]$, and assume that the graphon convolutions in all layers of this WNN have $K$ filter taps [cf. (6)]. Let $\mathbf{\Phi}(\mathbf{x}_n;\mathcal{H},\mathbf{S}_n)$ be a GNN sampled from $\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})$ as in (9). Under assumptions (1),(2),(3), and (5) with probability $1-\xi$ it holds that,*

$$\|\nabla_{\mathcal{H}}\mathbf{\Phi}(X;\mathcal{H},\mathbf{W}) - \nabla_{\mathcal{H}}\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n)\| \leq \sqrt{KF^{L-1}}\Bigg(2L^2F^{2L-2}\bigg(1+\frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}}\bigg)\frac{2\Big(1+\sqrt{n\log(\frac{2n}{\xi})}\Big)}{n} + \frac{2F^{L-1}L}{n} + 8L^2F^{2L-2}c\Bigg).$$

*Proof.* We will first show that the gradient with respect to any arbitrary element $[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger} \in \mathbb{R}$ of $\mathcal{H}$ can be uniformly bounded. Note that the maximum is attained if $l^\dagger = 1$. Without loss of generality, assuming $l^\dagger > l-1$, we can begin by using the definition given in equation (8) of the output of the WNN as follows,

$$\begin{aligned}
&\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X;\mathcal{H},\mathbf{W}) - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n)\| \\
&= \|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X^f_L - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X^f_{nL}\| \qquad (57)\\
&= \Bigg\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\rho\left(\sum_{g=1}^{F_{l-1}}\sum_{k=1}^{K-1}(T^{(k)}_{\mathbf{W}}X^g_{l-1})[\mathbf{H}_{lk}]_{gf}\right) \\
&\quad - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\rho\left(\sum_{g=1}^{F_{l-1}}\sum_{k=1}^{K-1}(T^{(k)}_{\mathbf{W}_n}X^g_{nl-1})[\mathbf{H}_{lk}]_{gf}\right)\Bigg\|. \qquad (58)
\end{aligned}$$

Taking derivatives by applying the chain rule, and applying the triangle inequality it yields,

$$\begin{aligned}
&\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X^f_L - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X^f_{nL}\| \\
&\leq \Bigg\|\Bigg(\nabla\rho\bigg(\sum_{g=1}^{F_{l-1}}\sum_{k=1}^{K-1}(T^{(k)}_{\mathbf{W}}X^g_{l-1})[\mathbf{H}_{lk}]_{gf}\bigg) - \nabla\rho\bigg(\sum_{g=1}^{F_{l-1}}\sum_{k=1}^{K-1}(T^{(k)}_{\mathbf{W}_n}X^g_{nl-1})[\mathbf{H}_{lk}]_{gf}\bigg)\Bigg) \\
&\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\bigg(\sum_{g=1}^{F_{l-1}}\sum_{k=1}^{K-1}(T^{(k)}_{\mathbf{W}}X^g_{l-1})[\mathbf{H}_{lk}]_{gf}\bigg)\Bigg\| \qquad (59)\\
&+ \Bigg\|\nabla\rho\bigg(\sum_{g=1}^{F_{l-1}}\sum_{k=1}^{K-1}(T^{(k)}_{\mathbf{W}_n}X^g_{nl-1})[\mathbf{H}_{lk}]_{gf}\bigg) \qquad (60)\\
&\Bigg(\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\sum_{g=1}^{F_{l-1}}\sum_{k=1}^{K-1}(T^{(k)}_{\mathbf{W}}X^g_{l-1})[\mathbf{H}_{lk}]_{gf} - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\sum_{g=1}^{F_{l-1}}\sum_{k=1}^{K-1}(T^{(k)}_{\mathbf{W}_n}X^g_{nl-1})[\mathbf{H}_{lk}]_{gf}\Bigg)\Bigg\|.
\end{aligned}$$

We can now use Cauchy-Schwartz inequality, Assumptions 3, 4, and Proposition 3 to bound the terms regarding the gradient of the non-linearity $\rho$, the loss function $\ell$, and the WNN respectively, as follows,

$$\begin{aligned}
&\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X^f_L - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X^f_{nL}\| \qquad (61)\\
&\leq \Bigg\|\sum_{g=1}^{F_{l-1}}\sum_{k=1}^{K-1}(T^{(k)}_{\mathbf{W}}X^g_{l-1})[\mathbf{H}_{lk}]_{gf} - \sum_{g=1}^{F_{l-1}}\sum_{k=1}^{K-1}(T^{(k)}_{\mathbf{W}_n}X^g_{nl-1})[\mathbf{H}_{lk}]_{gf}\Bigg\|F^{L-1}\|X_0\| \\
&+ \Bigg\|\sum_{g=1}^{F_{l-1}}\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\sum_{k=1}^{K-1}\Big((T^{(k)}_{\mathbf{W}}X^g_{l-1})[\mathbf{H}_{lk}]_{gf} - (T^{(k)}_{\mathbf{W}_n}X^g_{nl-1})[\mathbf{H}_{lk}]_{gf}\Big)\Bigg\|.
\end{aligned}$$

We can now apply the triangle inequality on the second term of the previous bound to obtain,

$$\begin{aligned}
&\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X_L^f - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X_{nL}^f\| \qquad (62)\\
&\leq \left\| \sum_{g=1}^{F_{l-1}} \sum_{k=1}^{K-1} (T_{\mathbf{W}}^{(k)} X_{l-1}^g)[\mathbf{H}_{lk}]_{gf} - \sum_{g=1}^{F_{l-1}} \sum_{k=1}^{K-1} (T_{\mathbf{W}_n}^{(k)} X_{nl-1}^g)[\mathbf{H}_{lk}]_{gf} \right\| F^{L-1} \|X_0\| \\
&+ \left\| \sum_{g=1}^{F_{l-1}} \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \sum_{k=1}^{K-1} \left( (T_{\mathbf{W}}^{(k)})[\mathbf{H}_{lk}]_{gf} - (T_{\mathbf{W}_n}^{(k)})[\mathbf{H}_{lk}]_{gf} \right) X_{nl-1}^g \right\| \\
&+ \sum_{g=1}^{F_{l-1}} \left\| \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \sum_{k=1}^{K-1} T_{\mathbf{W}_n}^{(k)} \left( X_{l-1}^g - X_{nl-1}^g \right) [\mathbf{H}_{lk}]_{gf}) \right\|.
\end{aligned}$$

Now note that as we are considering the case in which $l_\dagger < l-1$, using Cauchy-Schwartz inequality, we can use the same bound for the first and second term of the right hand side of the previous inequality. Since filters are non-expansive by Assumption 3, it yields

$$\begin{aligned}
&\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X_L^f - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X_{nL}^f\| \qquad (63)\\
&\leq 2 \left\| \sum_{g=1}^{F_{l-1}} \sum_{k=1}^{K-1} (T_{\mathbf{W}}^{(k)} X_{l-1}^g)[\mathbf{H}_{lk}]_{gf} - \sum_{g=1}^{F_{l-1}} \sum_{k=1}^{K-1} (T_{\mathbf{W}_n}^{(k)} X_{nl-1}^g)[\mathbf{H}_{lk}]_{gf} \right\| F^{L-1} \|X_0\| \\
&+ \sum_{g=1}^{F_{l-1}} \left\| \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \left( X_{l-1}^g - X_{nl-1}^g \right) \right\|.
\end{aligned}$$

Now notice, that the only term that remains to bound is the exact same bound we obtained in equation (57), but on the previous layer $L-2$. Hence, we conclude that by applying the same steps $L-2$ times, as the WNN has $L$ layers, we will obtain a bound for any element $[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}$ of tensor $\mathcal{H}$.

$$\begin{aligned}
&\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X_L^f - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X_{nL}^f\| \qquad (64)\\
&\leq 2LF^{L-2} \left\| \sum_{g=1}^{F_{l-1}} \sum_{k=1}^{K-1} (T_{\mathbf{W}}^{(k)} X_{l-1}^g)[\mathbf{H}_{lk}]_{gf} - \sum_{g=1}^{F_{l-1}} \sum_{k=1}^{K-1} (T_{\mathbf{W}_n}^{(k)} X_{nl-1}^g)[\mathbf{H}_{lk}]_{gf} \right\| F^{L-1} \|X_0\| \\
&+ \sum_{g=1}^{F_{l-1}} \left\| \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} \left( X_1^g - X_1^g \right) \right\|.
\end{aligned}$$

Note that the derivative of a convolutional filter $T_{\mathbf{H}}$ at coefficient $k^\dagger = i$, is itself a convolutional filter with coefficients $\mathbf{h}_i$ [cf. Definition 6]. The values of $\mathbf{h}_i$ are $[\mathbf{h}_i]_j = 1$ if $j = i$ and 0 otherwise. As $\mathbf{h}_i$ is itself a filter that verifies Assumption 2, as graphons are normalized. Thus, considering $l^\dagger = 0$, and using Propositions 1, 2, (Chung & Radcliffe, 2011, Theorem 1) and the triangle inequality, we obtain,

$$\left\| \mathbf{h}_{i*\mathbf{W}_n} X_{n0} - \mathbf{h}_{i*\mathbf{W}} X_0 \right\| \leq \left( \|\mathbf{W} - \mathbb{W}_n\| + \|\mathbb{W}_n - \mathbf{W}_n\| \right) \|X_0\| + \|X_{n0} - X_0\| \qquad (65)$$

$$\leq \left( 1 + \frac{\pi B_{\mathbf{W}_n}^c}{\delta_{\mathbf{W}\mathbf{W}_n}^c} \right) \frac{2\left( 1 + \sqrt{n \log(\frac{2n}{\xi})} \right)}{n} + \frac{1}{n} \qquad (66)$$

with probability $1-\xi$. In the previous expression, $\mathbb{W}_n$ is the template graphon [cf. Definition 2]. Now, substituting (64) into (65), and using Lemma 1, with probability $1-\xi$, it holds that,

$$\begin{aligned}
\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X_L^f - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}} X_{nL}^f\| \leq & 2L^2 F^{2L-2} \left( 1 + \frac{\pi B_{\mathbf{W}_n}^c}{\delta_{\mathbf{W}\mathbf{W}_n}^c} \right) \frac{2\left( 1 + \sqrt{n \log(\frac{2n}{\xi})} \right)}{n} \\
&+ \frac{2F^{L-1}L}{n} + 8L^2 F^{2L-2} c. \qquad (67)
\end{aligned}$$

To achieve the final result, note that tensor $\mathcal{H}$ has $KF^{L-1}$ elements, and each individual gradient is upper bounded by (67). □

**Lemma 3.** *Let $\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})$ be a WNN with $F_0 = F_L = 1$, and $F_l = F$ for $1 \leq l \leq L-1$. Let $c \in (0,1]$, and assume that the graphon convolutions in all layers of this WNN have $K$ filter taps [cf. (6)]. Let $\mathbf{\Phi}(\mathbf{x}_n;\mathcal{H},\mathbf{S}_n)$ be a GNN sampled from $\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})$ as in (9). Under Assumptions (1)–(5) with probability $1-\xi$ it holds that,*

$$\begin{aligned}\|\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})) - \nabla_{\mathcal{H}}\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n))\|& \\ \leq \sqrt{KF^{L-1}}\Bigg(3L^2F^{2L-2}\bigg(1+\frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}}\bigg)\frac{2\bigg(1+\sqrt{n\log(\frac{2n}{\xi})}\bigg)}{n}& \\ + \frac{4F^{L-1}L}{n} + 12L^2F^{2L-2}c\Bigg)&\end{aligned}$$

*Proof.* In order to analyze the norm of the gradient with respect to the tensor $\mathcal{H}$, we can start by taking the derivative with respect to a single element of the tensor, $[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}$. By deriving the loss function $\ell$ using the chain rule it yields,

$$\begin{aligned}\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\ell(Y,\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})) - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n))\|& \\ =\|\nabla\ell(Y,\mathbf{\Phi}(X;\mathcal{H},\mathbf{W}))\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})& \\ - \nabla\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n))\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n)\|.&\end{aligned} \tag{68}$$

By Cauchy-Schwartz, and the triangle inequality it holds,

$$\begin{aligned}&\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\ell(Y,\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})) - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n))\| \\ \leq&\|\nabla\ell(Y,\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})) - \nabla\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n))\|\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})\| \\ &+ \|\nabla\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n))\|\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X;\mathcal{H},\mathbf{W}) - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n)\|.\end{aligned} \tag{69}$$

By the triangle inequality and Assumption 4 it follows,

$$\begin{aligned}&\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\ell(Y,\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})) - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n))\| \\ &\leq \|\nabla\ell(Y,\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})) - \nabla\ell(Y,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n))\|\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})\| &(70)\\ &\|\nabla\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n)) - \nabla\ell(Y,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n))\|\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})\| &(71)\\ &+ \|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X;\mathcal{H},\mathbf{W}) - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n)\| \\ &\leq (\|Y_n - Y\| + \|\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n)) - \mathbf{\Phi}(X;\mathcal{H},\mathbf{W}))\|)\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})\| &(72)\\ &+ \|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X;\mathcal{H},\mathbf{W}) - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n)\|.\end{aligned}$$

Now we can use Lemmas 1–2, Propositions 1, and 3, and Assumption 1 to obtain,

$$\begin{aligned}\|\nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\ell(Y,\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})) - \nabla_{[\mathbf{H}_{l^\dagger k^\dagger}]_{g^\dagger f^\dagger}}\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n))\|& \\ \leq \bigg(3L^2F^{2L-2}\bigg(1+\frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}}\bigg)\frac{2\bigg(1+\sqrt{n\log(\frac{2n}{\xi})}\bigg)}{n}& \\ + \frac{4F^{L-1}L}{n} + 12L^2F^{2L-2}c\bigg)&\end{aligned} \tag{73}$$

Noting that tensor $\mathcal{H}$ has $KF^{L-1}$ elements, and each individual term can be bounded by (73), the desired result is attained. □

**Definition 7.** *We define the constant $\gamma$ as,*

$$\gamma = 12\sqrt{KF^{L-1}}L^2F^{2L-2}, \tag{74}$$

*where $K$ is the number of features, $L$ is the number of layers, and $K$ is the number of filter taps of the GNN.*

We will present a more comprehensive statement of Theorem 1, where we include all the smaller order terms in (15). Notice that the statement of Theorem 1 in the main body of the paper omits these terms in order to simplify the exposition of the main result. In practice, these smaller order terms vanish faster as $n$ increases.

**Theorem 1.** *Consider the ERM problem in* (12) *and let* $\mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})$ *be an* $L$*-layer WNN with* $F_0 = F_L = 1$*, and* $F_l = F$ *for* $1 \leq l \leq L-1$*. Let* $c \in (0, 1]$ *and assume that the graphon convolutions in all layers of this WNN have* $K$ *filter taps [cf.* (6)*]. Let* $\mathbf{\Phi}(\mathbf{x}_n; \mathcal{H}, \mathbf{S}_n)$ *be a GNN sampled from* $\mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})$ *as in* (9)*. Under assumptions AS1–AS5, it holds that*

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})) - \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}, \mathbf{W}_n))\|] \\
&\qquad \leq \sqrt{KF^{L-1}} \Bigg( 6L^2 F^{2L-2} \left( 1 + \frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}} \right) \frac{\left( 1 + \sqrt{n \log(2n^{3/2})} \right)}{n} \\
&\qquad + \frac{4F^{L-1}L}{n} + 12L^2 F^{2L-2} c \Bigg) + \frac{2F^{2L}\sqrt{K}}{\sqrt{n}} \qquad (75)
\end{aligned}$$

*where* $Y_n$ *is the graphon signal induced by* $[\mathbf{y}_n]_i = Y(u_i)$*,* $u_i = (i-1)/n$ *for* $1 \leq i \leq n$ *[cf.* (10)*]. The fixed constants* $B^c_{\mathbf{W}}$ *and* $\delta^c_{\mathbf{W}\mathbf{W}_n}$ *are the* $c$*-band cardinality and the* $c$*-eigenvalue margin of* $\mathbf{W}$ *and* $\mathbf{W}_n$ *respectively [cf. Definitions 4,5 in the supplementary material].*

*Proof of Theorem 1.* We begin by considering the event $A_n$ such that,

$$\begin{aligned}
A_n = \Bigg( &\|\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})) - \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}, \mathbf{W}_n))\| \qquad (76) \\
&\leq \sqrt{KF^{L-1}} \Bigg( 3L^2 F^{2L-2} \left( 1 + \frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}} \right) \frac{2\left( 1 + \sqrt{n \log(\frac{2n}{\xi})} \right)}{n} \\
&+ \frac{4F^{L-1}L}{n} + 12L^2 F^{2L-2} c \Bigg) \Bigg).
\end{aligned}$$

Thus, by considering the disjoint events $A_n$, and $A_n^c$, and denoting $\mathbf{1}(\cdot)$ the indicator function, the expectation can be separated as follows,

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})) - \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}, \mathbf{W}_n))\|] \\
&= \mathbb{E}[\|\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})) - \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}, \mathbf{W}_n))\| \mathbf{1}(A_n)] \\
&\quad + \mathbb{E}[\|\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})) - \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}, \mathbf{W}_n))\| \mathbf{1}(A_n^c)] \qquad (77)
\end{aligned}$$

We can bound the term regarding $A_n^c$ using the chain rule, Cauchy-Schwartz inequality, Assumption 4, and Proposition 3 as follows,

$$\begin{aligned}
&\|\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})) - \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}, \mathbf{W}_n))\| \\
&\leq \|\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W}))\| + \|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}, \mathbf{W}_n))\| \qquad (78) \\
&\leq \|\nabla \ell(Y, \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W}))\| \|\nabla_{\mathcal{H}} \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})\| \\
&\quad + \|\nabla \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}, \mathbf{W}_n))\| \|\nabla_{\mathcal{H}} \mathbf{\Phi}(X_n; \mathcal{H}, \mathbf{W}_n)\| \qquad (79) \\
&\leq \|\nabla_{\mathcal{H}} \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})\| + \|\nabla_{\mathcal{H}} \mathbf{\Phi}(X_n; \mathcal{H}, \mathbf{W}_n)\| \qquad (80) \\
&\leq 2F^{2L}\sqrt{K} \qquad (81)
\end{aligned}$$

Returning to equation (77), we can substitute the bound obtained in equation (81), and by taking $P(A_n) = 1 - \xi$, and using Lemma 3, it yields,

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})) - \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}, \mathbf{W}_n))\|] \\
&\leq (1-\xi)\sqrt{KF^{L-1}} \Bigg( 3L^2 F^{2L-2} \left( 1 + \frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}} \right) \frac{2\left( 1 + \sqrt{n \log(\frac{2n}{\xi})} \right)}{n} \\
&+ \frac{4F^{L-1}L}{n} + 12L^2 F^{2L-2} c \Bigg) + \xi 2F^{2L}\sqrt{K} \qquad (82)
\end{aligned}$$

To complete the proof, set $\xi = \frac{1}{\sqrt{n}}$. □

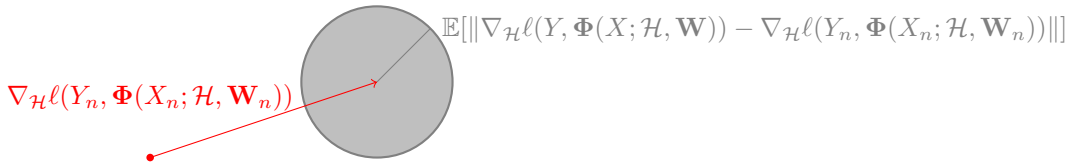


Figure 3: In order to satisfy the property that the inner product between the gradient on the GNN $\nabla_{\mathcal{H}}\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H},\mathbf{W}_n))$ and the gradient on the graphon $\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{H},\mathbf{W}))$ is positive, we rely on the condition provided in Theorem 2.

## D PROOF OF THEOREM 2

**Definition 8** (Stopping time)**.** *We define the stopping time $k^*$ as,*

$$k^* = \min_{k\geq 0}\{\|\nabla_{\mathcal{H}}\mathbf{\Phi}(X;\mathcal{H}_k,\mathbf{W}_n)\| \leq \sqrt{KF^{L-1}}12L^2F^{2L-2}c\}. \tag{83}$$

**Definition 9** (Constant $\psi$)**.**

**Lemma 4.** *Under Assumptions 4, 5, and 6, the gradient of the loss function $\ell$ with respect to the parameters of the GNN $\mathcal{H}$ is $A_{\nabla\ell}$-Lipschitz,*

$$\|\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{A},\mathbf{W}))-\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{B},\mathbf{W}))\| \leq A_{\nabla\ell}\|\mathcal{A}-\mathcal{B}\| \tag{84}$$

*where $A_{\nabla\ell} = (A_{\nabla\Phi} + A_{\mathbf{\Phi}}F^{2L}\sqrt{K})$.*

*Proof.* To begin with, we can apply the chain rule to obtain,

$$\begin{aligned}&\|\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{A},\mathbf{W}))-\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{B},\mathbf{W}))\|\\ &=\|\nabla\ell(Y,\mathbf{\Phi}(X;\mathcal{A},\mathbf{W}))\nabla_{\mathcal{H}}\mathbf{\Phi}(X;\mathcal{A},\mathbf{W})-\nabla\ell(Y,\mathbf{\Phi}(X;\mathcal{B},\mathbf{W}))\nabla_{\mathcal{H}}\mathbf{\Phi}(X;\mathcal{B},\mathbf{W})\|\end{aligned} \tag{85}$$

By applying the triangle inequality, and Cauchy-Schwartz it yields,

$$\begin{aligned}&\|\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{A},\mathbf{W}))-\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{B},\mathbf{W}))\|\\ &\leq\|\nabla\ell(Y,\mathbf{\Phi}(X;\mathcal{A},\mathbf{W}))\|\|\nabla_{\mathcal{H}}\mathbf{\Phi}(X;\mathcal{A},\mathbf{W})-\nabla_{\mathcal{H}}\mathbf{\Phi}(X;\mathcal{B},\mathbf{W})\|\\ &+\|\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{A},\mathbf{W}))-\ell(Y,\mathbf{\Phi}(X;\mathcal{B},\mathbf{W}))\|\|\nabla_{\mathcal{H}}\mathbf{\Phi}(X;\mathcal{B},\mathbf{W})\|\end{aligned} \tag{86}$$

We can now use Assumptions 1, 4, 5, and 6 as well as Proposition 3, to obtain

$$\|\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{A},\mathbf{W}))-\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{B},\mathbf{W}))\| \leq (A_{\nabla\Phi}+A_{\mathbf{\Phi}}F^{2L}\sqrt{K}\|\mathcal{A}-\mathcal{B}\|) \tag{87}$$

Completing the proof. □

**Lemma 5.** *Consider the ERM problem in (12) and let $\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})$ be an $L$-layer WNN with $F_0 = F_L = 1$, and $F_l = F$ for $1 \leq l \leq L-1$. Let $c \in (0,1]$ and assume that the graphon convolutions in all layers of this WNN have $K$ filter taps [cf. (6)]. Let $\mathbf{\Phi}(\mathbf{x}_n;\mathcal{H},\mathbf{S}_n)$ be a GNN sampled from $\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})$ as in (9). Under assumptions AS1–AS6, let the following condition be satisfied for every $k$,*

$$\begin{aligned}&\sqrt{KL^{L-1}}\left(6L^2F^{2L-2}\left(1+\frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}}\right)\frac{\left(1+\sqrt{n\log(2n^{3/2})}\right)}{n}+\frac{4F^{L-1}L}{n}+12L^2F^{2L-2}c\right)\\ &+\frac{2F^{2L}\sqrt{K}}{\sqrt{n}}<\frac{1-A_{\nabla\ell}\eta_k}{2}\|\nabla_{\mathcal{H}}\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H}_k,\mathbf{W}_n)\|.\end{aligned} \tag{88}$$

*then the first $k^*$ iterates generated by equation (14), $\mathbf{1}(k\leq k^*)\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_k,\mathbf{W}_n))$ form a positive super-martingale with respect to the filtration $\mathcal{F}_k$ generated by the history of the Algorithm up to step $k$ [i.e., $\{X,Y,X_n,Y_n,\mathbf{W}_n\}_k,\{X,Y,X_n,Y_n,\mathbf{W}_n\}_{k-1},\dots,\{X,Y,X_n,Y_n,\mathbf{W}_n\}_0$]. Where $k^*$ is the stopping time defined in Definition 8, and $\mathbf{1}(\cdot)$ is the indicator function.*

*Proof.* To begin with, $\mathbf{1}(k < k^*)\ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) \in \mathcal{F}_k$, where $\mathcal{F}_k$ is the filtration generated by the history of the Algorithm up to $k$. Note that the loss function $\ell$ is positive by Assumption 4. It remains to be shown the inequality expression of the super-martingale. For $k > k^*$, the inequality is trivially verified as the indicator function $\mathbf{1}(k \leq k^*) = 0$ for $k > k^*$. For $k \leq k^*$, as in Bertsekas & Tsitsiklis (2000), we define a continuous function $g(\epsilon)$ that takes the value of the loss function on the Graphon data on iteration $k+1$ at $\epsilon = 1$, and on iteration $k$ on $\epsilon = 0$ as follows,

$$g(\epsilon) = \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k - \epsilon\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)), \mathbf{W})). \tag{89}$$

Note that function $g(\epsilon)$, is evaluated on the graphon data $Y, X, \mathbf{W}$, but the steps are controlled by the induced graphon data $Y_n, X_n, \mathbf{W}_n$. Applying the chain rule, the derivative of $g(\epsilon)$ with respect to $\epsilon$ can be obtain as follows,

$$\begin{aligned} &\frac{\partial}{\partial \epsilon} g(\epsilon) = \\ &- \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k - \epsilon\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(\mathcal{H}_k; \mathbf{W}_n; X_n)), \mathbf{W})) \eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k; \mathbf{W}_n)). \end{aligned} \tag{90}$$

Now note that the difference in the loss function $\ell$ between iterations $k+1$ and $k$ can be written as the difference between $g(\epsilon = 1)$ and $g(\epsilon = 0)$ as follows,

$$g(1) - g(0) = \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_{k+1}, \mathbf{W})) - \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})). \tag{91}$$

Computing the integration of the derivative of $g(\epsilon)$ between $[0, 1]$ it yields

$$\ell(Y, \mathbf{\Phi}(X; \mathcal{H}_{k+1}, \mathbf{W})) - \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) = g(1) - g(0) = \int_0^1 \frac{\partial}{\partial \epsilon} g(\epsilon) d\epsilon \tag{92}$$

$$\begin{aligned} &= \int_0^1 \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k - \epsilon\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)), \mathbf{W})) \\ &\quad (-)\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)) d\epsilon. \end{aligned} \tag{93}$$

Note that the last term of the previous integral does not depend on $\epsilon$. Besides, we can sum and subtract $\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(\mathcal{H}_k, \mathbf{W}, X))$ inside the integral, to obtain,

$$\begin{aligned} &\ell(Y, \mathbf{\Phi}(X; \mathcal{H}_{k+1}, \mathbf{W})) - \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) \\ &= (-)\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)) \\ &\int_0^1 \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k - \epsilon\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)), \mathbf{W})) \\ &\qquad + \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) - \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) d\epsilon \end{aligned} \tag{94}$$

$$\begin{aligned} &= -\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)) \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) \int_0^1 d\epsilon \\ &- \eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)) \int_0^1 \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(\mathcal{H}_k - \epsilon\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)), \mathbf{W}, X)) \\ &- \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) d\epsilon. \end{aligned} \tag{95}$$

We can now apply the Cauchy-Schwartz inequality to the last term on the previous inequality, and take the norm of the integral, which is smaller that the integral of the norm to obtain,

$$\begin{aligned} &\ell(Y, \mathbf{\Phi}(X; \mathcal{H}_{k+1}, \mathbf{W})) - \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) \\ &\leq -\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(\mathcal{H}_k; \mathbf{W}_n; X_n)) \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(\mathcal{H}_k, \mathbf{W}, X)) \\ &+ \eta_k \|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(\mathcal{H}_k; \mathbf{W}_n; X_n))\| \int_0^1 \Big\| \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(\mathcal{H}_k - \epsilon\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(\mathcal{H}_k; \mathbf{W}_n; X_n)), \mathbf{W}, X)) \\ &- \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(\mathcal{H}_k, \mathbf{W}, X)) \Big\| d\epsilon. \end{aligned} \tag{96}$$

Under Lemma 4, we can take the Lipschitz bound on the gradient on the loss function with respect to the parameters, using $A_{\nabla\ell}$, to obtain,

$$\begin{aligned}
&\ell(Y, \mathbf{\Phi}(X; \mathcal{H}_{k+1}, \mathbf{W})) - \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_{k}, \mathbf{W})) \\
&\leq -\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)) \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) \\
&\quad + A_{\nabla\ell} \eta_k \|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n))| \int_0^1 \left\| \eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)) \right\| \epsilon d\epsilon \qquad (97) \\
&\leq -\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)) \nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) \\
&\quad + \frac{\eta_k^2 A_{\nabla\ell}}{2} \|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n))\|^2. \qquad (98)
\end{aligned}$$

Instead of evaluating the internal product between the gradient on the graphon, and induced graphon, we will use Theorem 1, to bound their expected difference (cf. Figure 3 for intuition). We can add and subtract the gradient of the loss function on the induced graphon $\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(\mathcal{H}_k; \mathbf{W}_n; X_n))$, and use the Cauchy-Schwartz inequality to obtain,

$$\begin{aligned}
&\ell(Y, \mathbf{\Phi}(X; \mathcal{H}_{k+1}, \mathbf{W})) - \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_{k}, \mathbf{W})) \\
&\leq -\eta_k \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)) \\
&(\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) + \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)) - \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n))) \\
&+ \frac{\eta_k^2 A_{\nabla\Phi}}{2} \|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n))\|^2 \qquad (99) \\
&\leq -\eta_k \|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n))\|^2 \\
&+ \eta_k \|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)\| \|\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) - \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n))\| \\
&+ \frac{\eta_k^2 A_{\nabla\Phi}}{2} \|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n))\|^2. \qquad (100)
\end{aligned}$$

We can rearrange the previous expression, to obtain,

$$\begin{aligned}
&\eta_k \|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)\|^2 \Bigg( 1 - \frac{A_{\nabla\ell} \eta_k}{2} \\
&- \frac{\|\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) - \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n))\|}{\|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)\|} \Bigg) \\
&\leq \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) - \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_{k+1}, \mathbf{W})). \qquad (101)
\end{aligned}$$

We can now take the conditional expectation with respect to the filtration $\mathcal{F}_n$ to obtain,

$$\begin{aligned}
&\mathbb{E}[\ell(Y, \mathbf{\Phi}(X; \mathcal{H}_{k+1}, \mathbf{W}))|\mathcal{F}_k] \\
&\leq \eta_k \|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)\|^2 \Bigg( 1 - \frac{A_{\nabla\ell} \eta_k}{2} \\
&- \mathbb{E}\left[ \frac{\|\nabla_{\mathcal{H}} \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})) - \nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n))\|}{\|\nabla_{\mathcal{H}} \ell(Y_n, \mathbf{\Phi}(X_n; \mathcal{H}_k, \mathbf{W}_n)\|} \middle| \mathcal{F}_k \right] \Bigg) \\
&\quad + \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})). \qquad (102)
\end{aligned}$$

As step size $\eta_k > 0$, and by definition norms are non-negative, using Theorem 1, as condition (88) holds for $k \leq k^*$, then

$$\mathbb{E}[\ell(Y, \mathbf{\Phi}(X; \mathcal{H}_{k+1}, \mathbf{W}))|\mathcal{F}_k] \leq \ell(Y, \mathbf{\Phi}(X; \mathcal{H}_k, \mathbf{W})). \qquad (103)$$

By definition of super-martingale as in Durrett (2019), we complete the proof. □

**Lemma 6.** *Consider the ERM problem in* (12) *and let* $\mathbf{\Phi}(X; \mathcal{H}, \mathbf{W})$ *be an L-layer WNN with* $F_0 = F_L = 1$*, and* $F_l = F$ *for* $1 \leq l \leq L-1$*. Let* $c \in (0, 1]$ *and assume that the graphon convolutions in all layers of this WNN have K filter taps [cf.* (6)*]. Let* $\mathbf{\Phi}(\mathbf{x}_n; \mathcal{H}, \mathbf{S}_n)$ *be a GNN*

*sampled from* $\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})$ *as in* (9)*. Under assumptions AS1–AS6, for any* $\epsilon \in (0, 1 - A_{\nabla\ell}\eta)$*, if the iterates generated by* (14)*, satisfy,*

$$\sqrt{KL^{L-1}}\left(6L^2F^{2L-2}\left(1+\frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}}\right)\frac{\left(1+\sqrt{n\log(2n^{3/2})}\right)}{n}+\frac{4F^{L-1}L}{n}+12L^2F^{2L-2}c\right)$$
$$+\frac{2F^{2L}\sqrt{K}}{\sqrt{n}}<\frac{1-A_{\nabla\ell}\eta_k-\epsilon}{2}\|\nabla_{\mathcal{H}}\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H}_k,\mathbf{W}_n)\|. \tag{104}$$

*then the expected value of the stopping time* $k^*$ *[cf. Definition 8], is finite, i.e.,*

$$\mathbb{E}[k^*]=\mathcal{O}(1/\epsilon) \tag{105}$$

*Proof.* Given the iterates at $k = k^*$, and the initial values at $k = 0$, we can express the expected difference between the loss $\ell$, as the summation over the difference of iterates as follows,

$$\mathbb{E}[\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_0,\mathbf{W}))-\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k^*},\mathbf{W}))]=$$
$$\mathbb{E}\left[\sum_{k=1}^{k^*}\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k-1},\mathbf{W}))-\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_k,\mathbf{W})\right] \tag{106}$$

Taking the expected value with respect to the final iterate $k = k^*$, we get,

$$\mathbb{E}\left[\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k^0}))-\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k^*}))\right]$$
$$=\mathbb{E}_{k^*}\left[\mathbb{E}\left[\sum_{k=1}^{k^*}\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k-1},\mathbf{W}))-\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_k,\mathbf{W})\right]\middle|k^*\right] \tag{107}$$
$$=\sum_{t=0}^{\infty}\mathbb{E}\left[\sum_{k=1}^{t}\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k-1},\mathbf{W}))-\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_k,\mathbf{W})\right]P(k^*=t) \tag{108}$$

Using condition (104), and Lemma 5 for any $k \leq k^*$, it verifies

$$\mathbb{E}\left[\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k-1},\mathbf{W}))-\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_k,\mathbf{W}))\right]\geq\eta(\sqrt{KF^{L-1}}12L^2F^{2L-2}c)^2\epsilon \tag{109}$$

Thus, coming back to (108),

$$\mathbb{E}\left[\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k^0},\mathbf{W}))-\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k^*},\mathbf{W}))\right]\geq\eta(\sqrt{KF^{L-1}}12L^2F^{2L-2}c)^2\epsilon\sum_{t=0}^{\infty}tP(k^*=t) \tag{110}$$
$$\geq\eta(\sqrt{KF^{L-1}}12L^2F^{2L-2}c)^2\epsilon\mathbb{E}[k^*] \tag{111}$$

Note that as the loss function $\ell$ is non-negative,

$$\frac{\mathbb{E}\left[\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k^0},\mathbf{W}))\right]}{\eta(\sqrt{KF^{L-1}}12L^2F^{2L-2}c)^2\epsilon}\geq\mathbb{E}[k^*] \tag{112}$$

Thus concluding that $k^* = \mathcal{O}(1/\epsilon)$. □

**Theorem 2.** *Consider the ERM problem in* (12) *and let* $\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})$ *be an L-layer WNN with* $F_0 = F_L = 1$*, and* $F_l = F$ *for* $1 \leq l \leq L-1$*. Let* $c \in (0, 1]$ *and assume that the graphon convolutions in all layers of this WNN have K filter taps [cf.* (6)*]. Let* $\mathbf{\Phi}(\mathbf{x}_n;\mathcal{H},\mathbf{S}_n)$ *be a GNN sampled from* $\mathbf{\Phi}(X;\mathcal{H},\mathbf{W})$ *as in* (9)*. Consider the iterates generated by equation* (14)*. Under*

*Assumptions AS1-AS6, for any fixed $\epsilon \in (0, 1 - A_{\nabla \ell}\eta)$, if at each step $k$ the number of nodes $n$ is picked such that it verifies*

$$\sqrt{KL^{L-1}}\left(6L^2F^{2L-2}\left(1+\frac{\pi B^c_{\mathbf{W}_n}}{\delta^c_{\mathbf{W}\mathbf{W}_n}}\right)\frac{\left(1+\sqrt{n\log(2n^{3/2})}\right)}{n}+\frac{4F^{L-1}L}{n}+12L^2F^{2L-2}c\right)$$
$$+\frac{2F^{2L}\sqrt{K}}{\sqrt{n}}<\frac{1-A_{\nabla\ell}\eta_k-\epsilon}{2}\|\nabla_{\mathcal{H}}\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H}_k,\mathbf{W}_n)\|. \tag{113}$$

*then in finite time we will achieve an iterate $k^*$ such that the coefficients $\mathcal{H}_{k^*}$ satisfy*

$$\mathbb{E}[\|\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k^*},\mathbf{W}))\|]\leq 24\sqrt{KF^{L-1}}L^2F^{2L-2}c \quad \textit{with probability } 1 \tag{114}$$

*where $A_{\nabla\ell\eta_k}=(A_{\nabla\Phi}+A_{\mathbf{\Phi}}F^{2L}\sqrt{K})$.*

*Proof.* We can use Lemma 6, to conclude that it must be the case that $P(k^*=\infty)=0$, which implies that, $P(k^*<\infty)=1$. Using stopping time $k^*$ condition [cf. Definition 8] and the triangle inequality, it yields,

$$\begin{aligned}\mathbb{E}[\|\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k^*},\mathbf{W}))\|]\leq&\|\nabla_{\mathcal{H}}\ell(Y_n,\mathbf{\Phi}(X_n;\mathcal{H}_{k^*},\mathbf{W}_n))\| \\ &+\mathbb{E}[\|\nabla_{\mathcal{H}}\ell(Y_n,\mathbf{\Phi}(X;\mathcal{H}_{k^*},\mathbf{W}_n))-\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X_n;\mathcal{H}_{k^*},\mathbf{W}_n))\|]\end{aligned} \tag{115}$$

Note that the iterates are constructed such that, for every $k$

$$\mathbb{E}[\|\nabla_{\mathcal{H}}\ell(Y_n,\mathbf{\Phi}(X;\mathcal{H}_k,\mathbf{W}_n))-\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X_n;\mathcal{H}_k,\mathbf{W}_n))\|]\leq\|\nabla_{\mathcal{H}}\ell(Y_n,\mathbf{\Phi}(X;\mathcal{H}_k,\mathbf{W}_n))\|. \tag{116}$$

Using the stopping time condition, the final result is attained as follows

$$\begin{aligned}\mathbb{E}[\|\nabla_{\mathcal{H}}\ell(Y,\mathbf{\Phi}(X;\mathcal{H}_{k^*},\mathbf{W}))\|]\leq&2\|\nabla_{\mathcal{H}}\ell(Y_n,\mathbf{\Phi}(X;\mathcal{H}_{k^*},\mathbf{W}_n))\| &(117)\\ \leq&24\sqrt{KF^{L-1}}L^2F^{2L-2}c. &(118)\end{aligned}$$

□